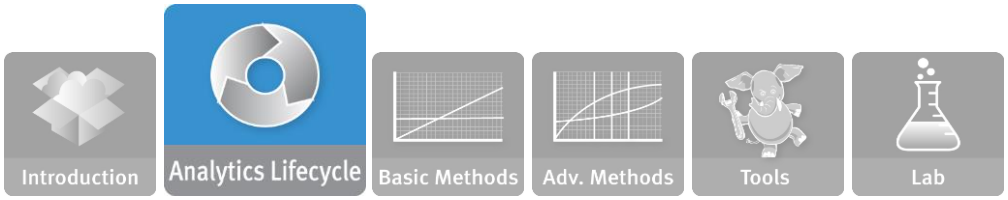
Information Systems

Introduction

Big Data Analytics

Presented by: Dr Sherin El Gokhy

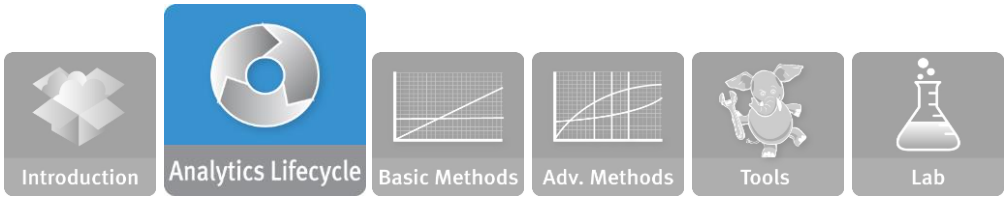
Module 2 – Data Analytics Lifecycle



Module 2: Data Analytics Lifecycle

Upon completion of this module, you should be able to:

- Apply the Data Analytics Lifecycle to a case study scenario
- Frame a business problem as an analytics problem
- Identify the four main deliverables (achievements) in an analytics project



Module 2: Data Analytics Lifecycle

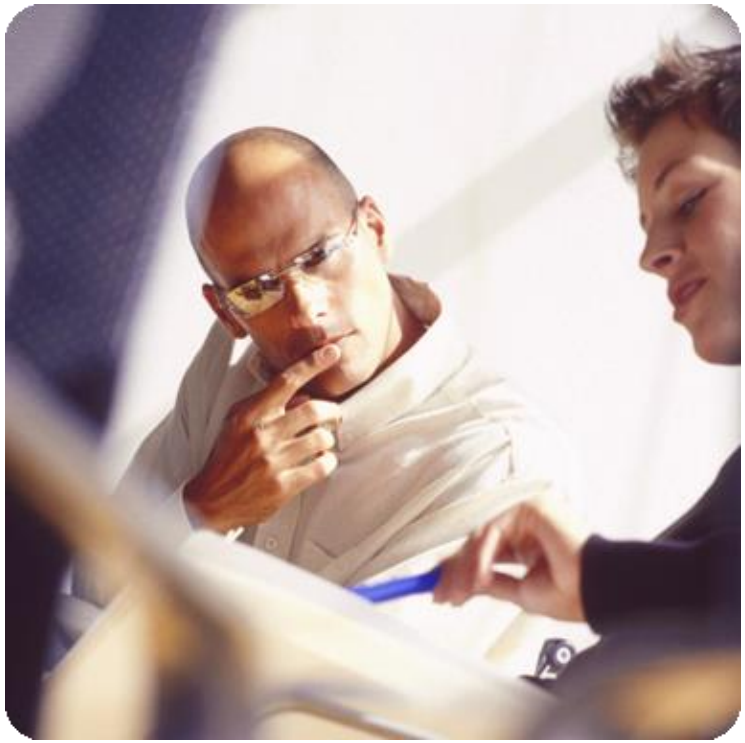
During this module the following topics are covered:

- Data Analytics Lifecycle
- Roles for a Successful Analytics Project
- Case Study to apply the data analytics lifecycle

How to Approach Your Analytics Problems

Your Thoughts?

- How do you currently approach your analytics problems?
- Do you follow a methodology or some kind of framework?
- How do you plan for an analytic project?



Value of Using the Data Analytics Lifecycle

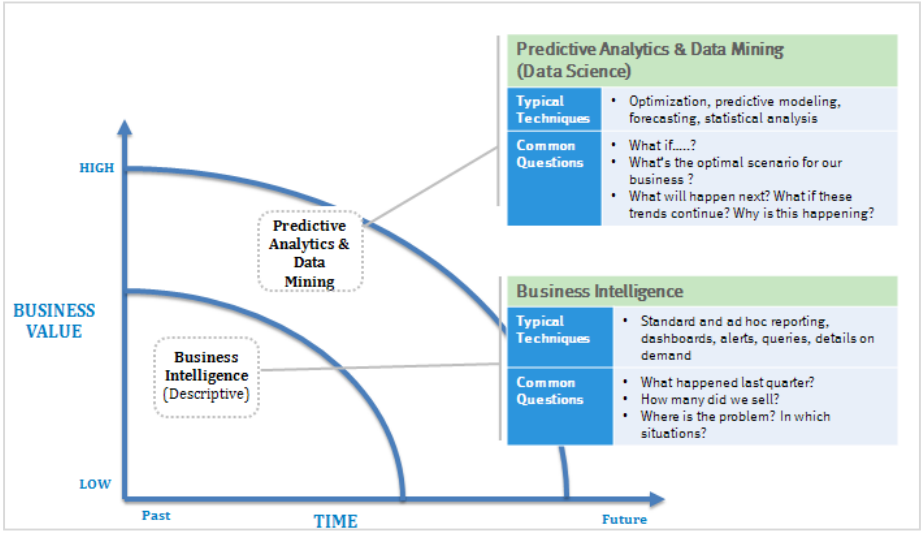
- Focus your time
- Ensure rigidity and completeness
- Enable better transition to members of the cross-functional analytic teams……. **Creating and documenting a process will help demonstrate rigor in your findings**
  - Repeatable
  - Scale to additional analysts
  - Support validity of findings

"A journey of a thousand miles begins with a single step" (Lao Tzu)

a well defined process enables you to break down complex problems into smaller steps

Need For a Process to Guide Data Science Projects

1. Well-defined processes can help guide any analytic project
2. Focus of Data Analytics Lifecycle is on Data Science projects, not business intelligence

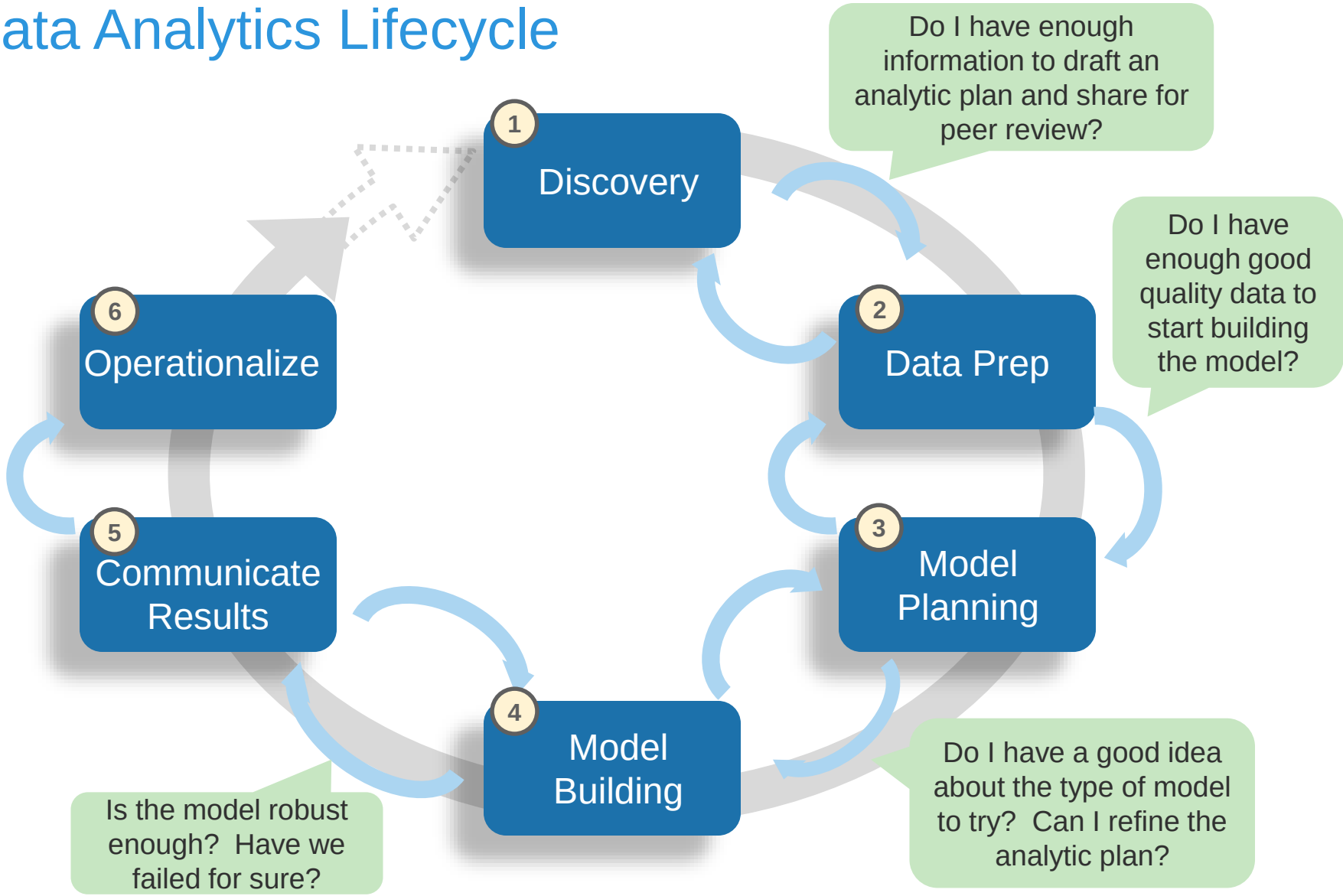


3. Data Science projects tend to require a more consultative approach, and differ from BI projects in a few ways
   - More due diligence in Discovery phase
   - More projects which lack shape or structure
   - Less predictable data

Key Roles for a Successful Analytic Project

Role	Description
Business User	Someone who **benefits from the end results** and **can consult and advise project team on value of end results and how these will be operationalized**
Project Sponsor	Person **responsible for the genesis of the project**, providing the motives for the project and core business problem, generally **provides the funding and will measure the degree of value from the final outputs of the working team**
Project Manager	**Ensure key objectives are met on time and at expected quality.**
Business Intelligence Analyst	Business domain expertise **with deep understanding of the data, KPIs, key metrics and business intelligence from a reporting perspective**
Data Engineer	**Deep technical skills to assist with tuning SQL queries for data management, extraction and support data realize to analytic sandbox**
Database Administrator (DBA)	Database Administrator who **provisions and configures database environment to support the analytical needs of the working team**
Data Scientist	Provide subject matter expertise for analytical techniques, **data modeling, applying valid analytical techniques to given business problems and ensuring overall analytical objectives are met**

Data Analytics Lifecycle

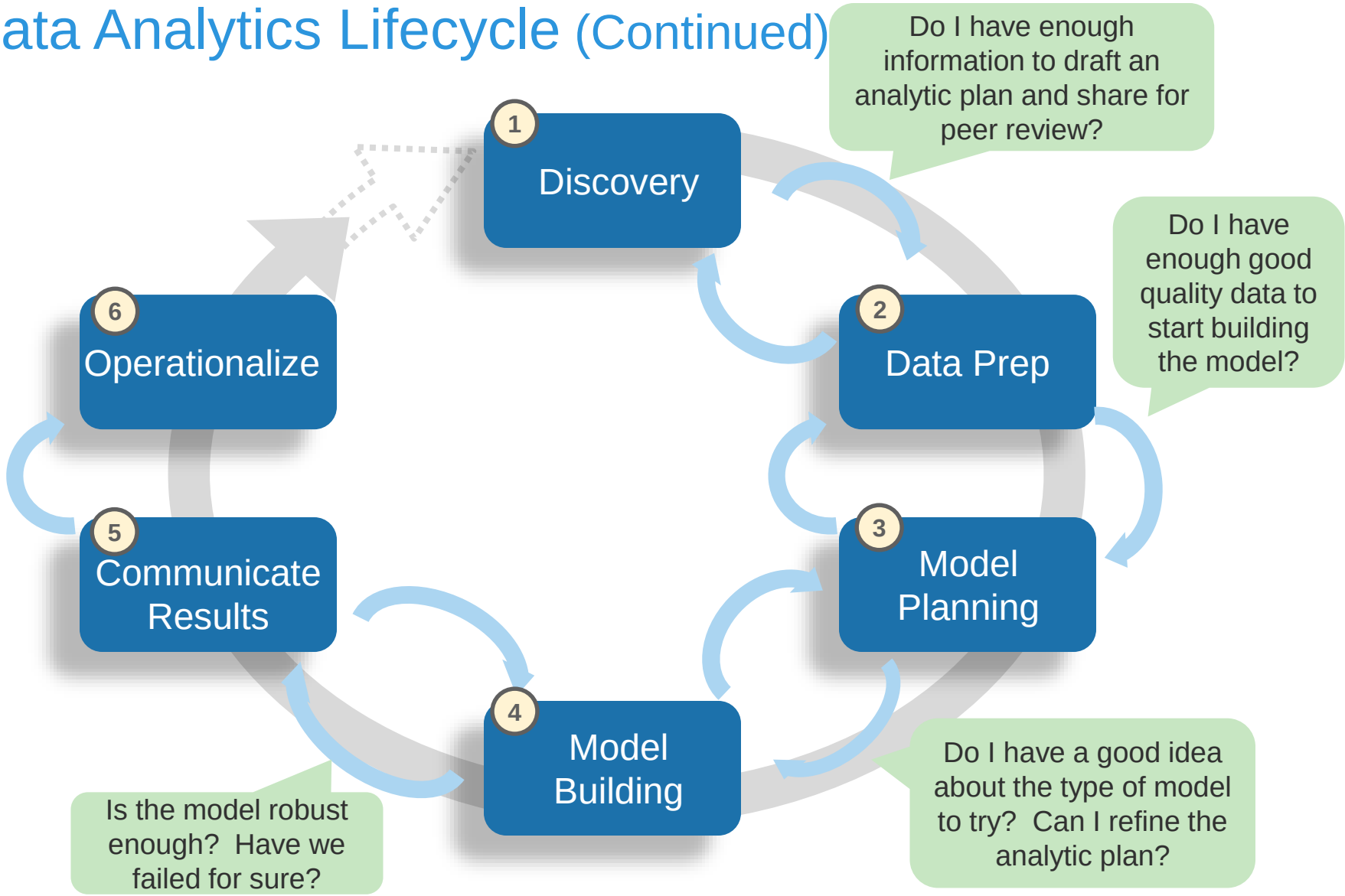


Data Analytics Lifecycle

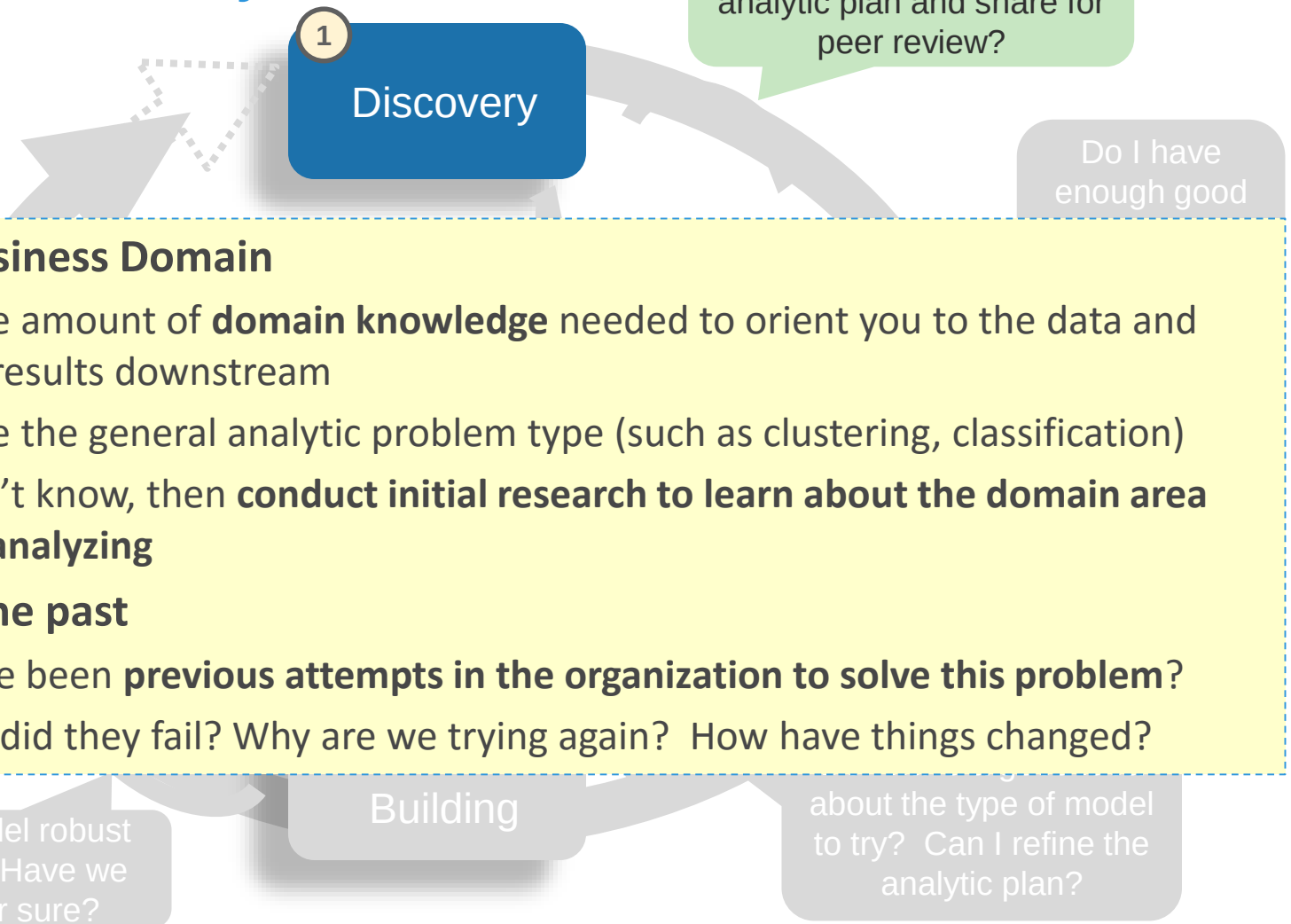
- As you can see in the graphic, **you can often learn something new in a phase to cause you to go back and refine work done in prior phases given new insights and information that you've uncovered**. For this reason, the graphic is shown as a cycle and the circular arrows are intended to convey that you can move iteratively between phases until you have sufficient information to continue moving forward.

- The green callouts represent questions to ask yourself **to gauge whether you have enough information and have made enough progress to move to the next phase of the process.**

Data Analytics Lifecycle (Continued)

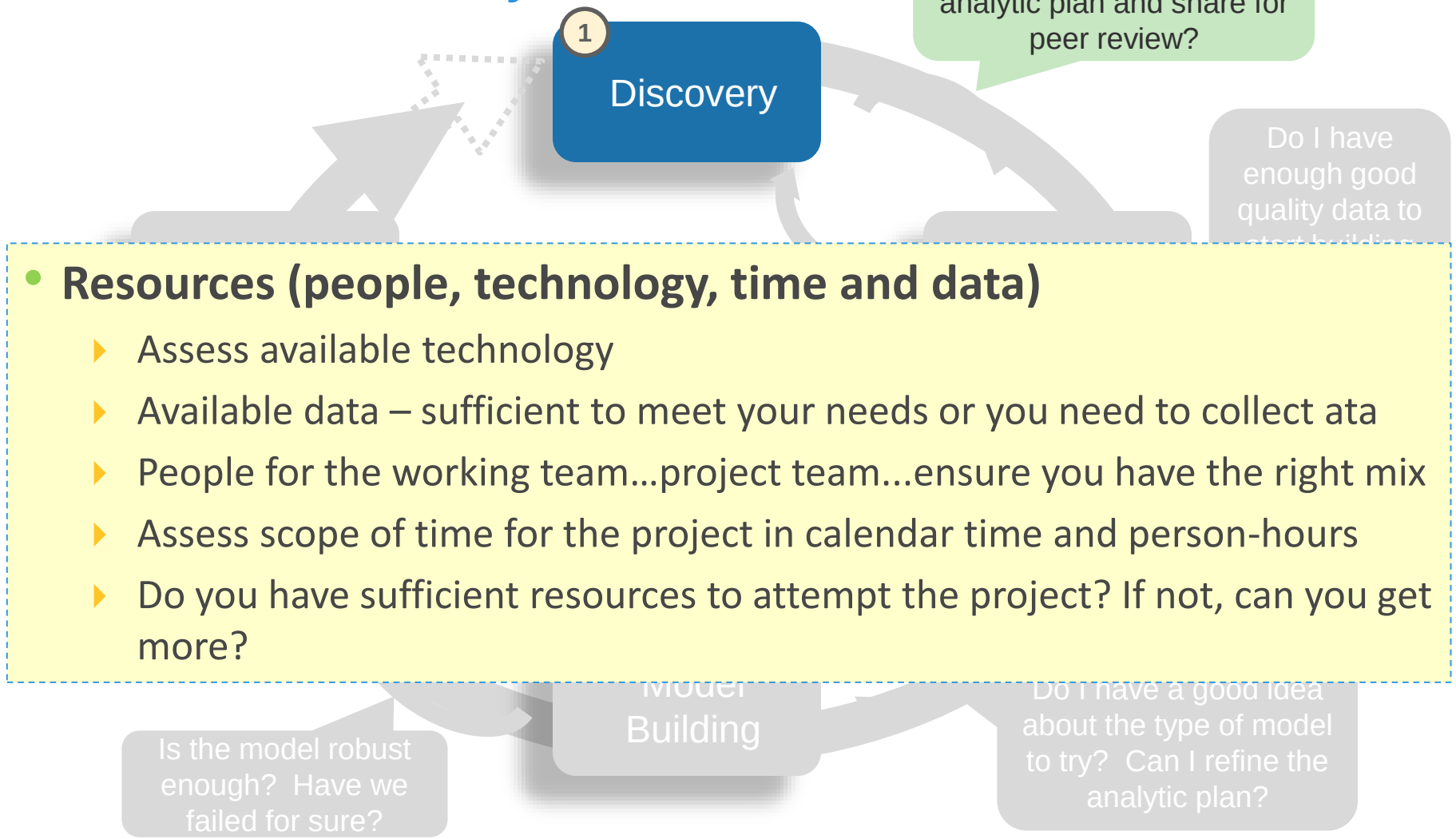


Data Analytics Lifecycle
Phase 1: Discovery



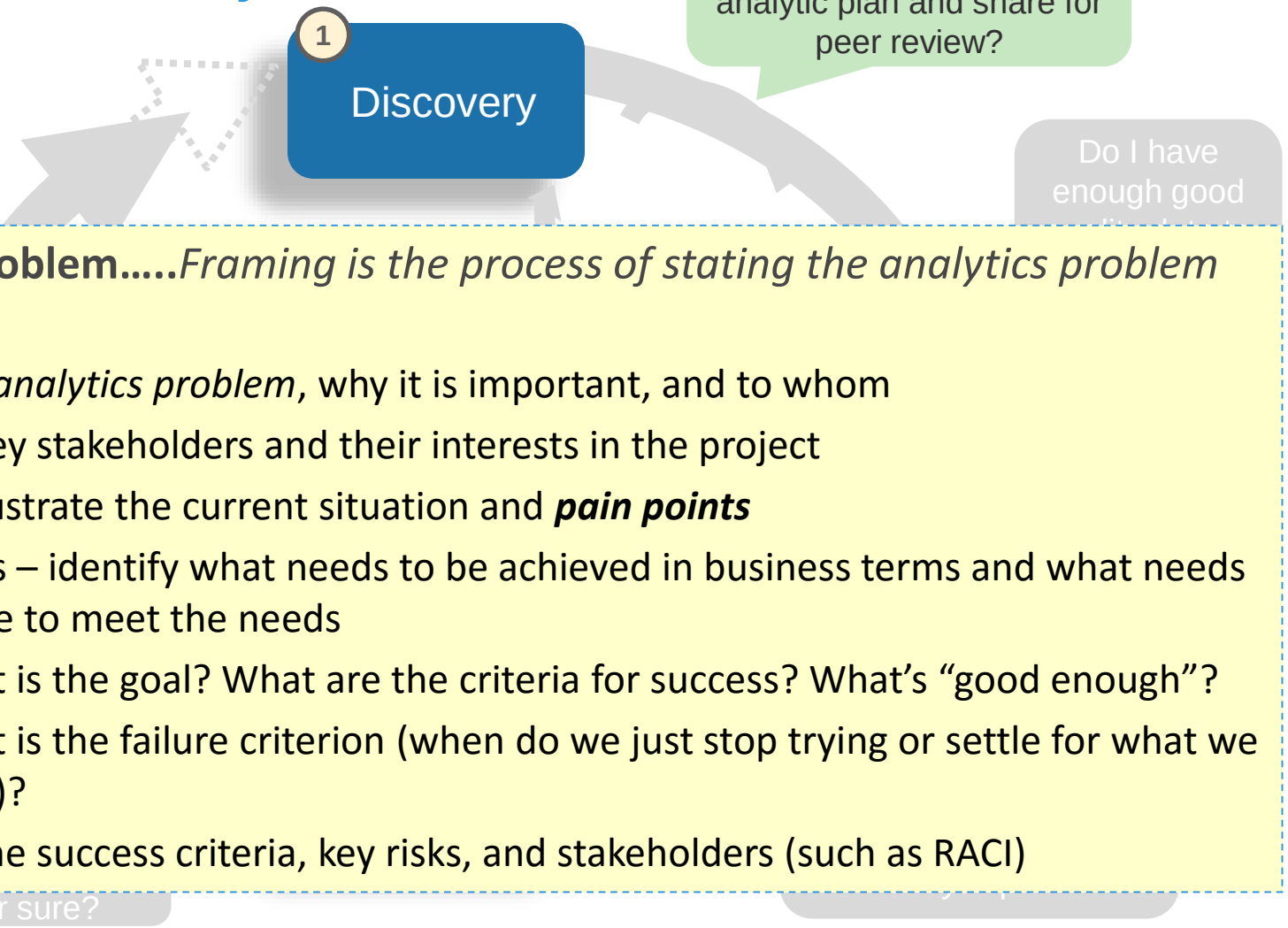
- **Learn the Business Domain**
  - Determine amount of **domain knowledge** needed to orient you to the data and interpret results downstream
  - Determine the general analytic problem type (such as clustering, classification)
  - If you don't know, then **conduct initial research to learn about the domain area you'll be analyzing**
- **Learn from the past**
  - Have there been **previous attempts in the organization to solve this problem**?
  - If so, why did they fail? Why are we trying again? How have things changed?

Data Analytics Lifecycle
Phase 1: Discovery



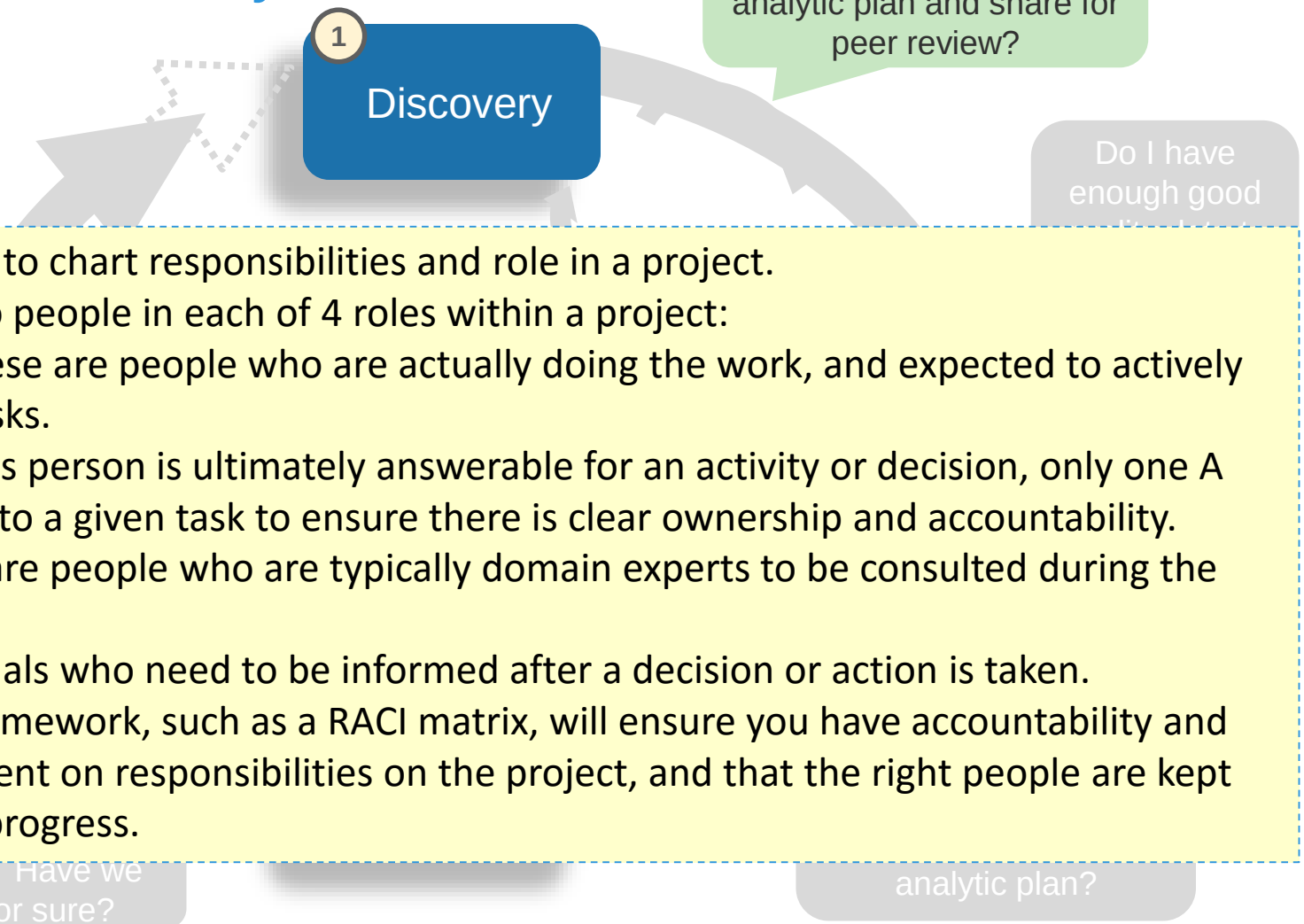
- **Resources (people, technology, time and data)**
  - Assess available technology
  - Available data – sufficient to meet your needs or you need to collect ata
  - People for the working team…project team...ensure you have the right mix
  - Assess scope of time for the project in calendar time and person-hours
  - Do you have sufficient resources to attempt the project? If not, can you get more?

Data Analytics Lifecycle
Phase 1: Discovery



- **Frame the problem…..***Framing is the process of stating the analytics problem to be solved*
  - *State the analytics problem*, why it is important, and to whom
  - Identify key stakeholders and their interests in the project
  - Clearly illustrate the current situation and ***pain points***
  - Objectives – identify what needs to be achieved in business terms and what needs to be done to meet the needs
    - What is the goal? What are the criteria for success? What's "good enough"?
    - What is the failure criterion (when do we just stop trying or settle for what we have)?
  - Identify the success criteria, key risks, and stakeholders (such as RACI)

Data Analytics Lifecycle
Phase 1: Discovery



- RACI is a way to chart responsibilities and role in a project.
- RACI refers to people in each of 4 roles within a project:

Responsible: these are people who are actually doing the work, and expected to actively complete the tasks.

Accountable: this person is ultimately answerable for an activity or decision, only one A can be assigned to a given task to ensure there is clear ownership and accountability.

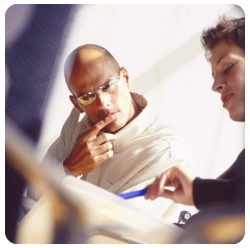
Consult: these are people who are typically domain experts to be consulted during the project.

Inform: individuals who need to be informed after a decision or action is taken.

- Creating a framework, such as a RACI matrix, will ensure you have accountability and clear agreement on responsibilities on the project, and that the right people are kept informed of progress.

Tips for Interviewing the Analytics Sponsor

- Even if you are "given" an analytic problem you should work with clients to clarify and frame the problem
  - You're typically handed solutions, you need to identify the problem and their desired outcome

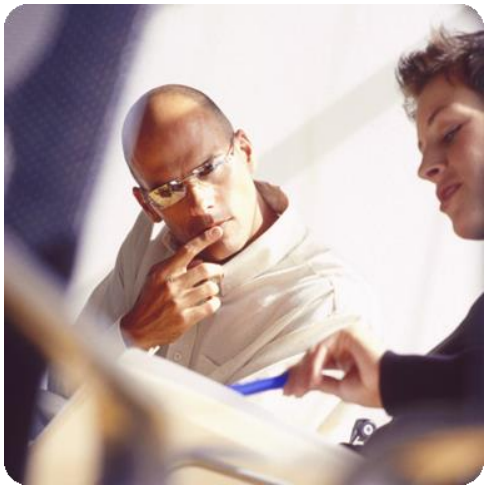


Sponsor Interview Tips

- Prepare for the interview – draft your questions, review with colleague, team
- Use open-ended questions, don't ask leading questions
- Probe for details, follow-up
- Don't fill every silence – give them time to think
- Let them express their ideas, don't put words in their mouth, let them share their feelings
- Ask clarifying questions, ask why – is that correct? Am I on target? Is there anything else?
- Use active listening – repeat it back to make sure you heard it correctly
- Don't express your opinions
- Be mindful of your body language and theirs – use eye contact, be attentive
- Minimize distractions
- Document what you heard and review it back with the sponsor

Tips for Interviewing the Analytics Sponsor Interview Questions

- What is the business problem you're trying to solve?
- What is your desired outcome?
- Will the focus and scope of the problem change if the following dimensions change:
  - Time – analyzing 1 year or 10 years worth of data?
  - People – how would this project change this?
  - Risk – conservative to aggressive
  - Resources – none to unlimited (tools, tech, …..)
  - Size and attributes of Data
- What data sources do you have?
- What industry issues may impact the analysis?
- What timelines are you up against?
- Who could provide insight into the project? Consulted?
- Who has final say on the project?



Data Analytics Lifecycle
Phase 1: Discovery

Do I have enough information to draft an analytic plan and share for peer review?

1 Discovery

Do I have enough good quality data to start building the model?

ood idea
of model
efine the
lan?

- **Formulate Initial Hypotheses**
  - IH, H_1 , H_2, H_3, … H_n
  - Gather and assess hypotheses from stakeholders and domain experts
  - Preliminary data exploration to inform discussions with stakeholders during the hypothesis forming stage
- **Identify Data Sources – Begin Learning the Data**
  - Aggregate sources for previewing the data and provide high-level understanding
  - Review the raw data
  - Determine the structures and tools needed
  - Scope the kind of data needed for this kind of problem

Using a Sample Case Study to Track the Phases in the Data Analytics Lifecycle

Mini Case Study: Churn Prediction for Yoyodyne Bank

Situation Synopsis

- Retail Bank, Yoyodyne Bank wants to improve the Net Present Value (NPV) and retention rate of customers
- They want to establish an effective marketing campaign targeting customers to reduce the churn rate by at least five percent
- The bank wants to determine whether those customers are worth retaining. In addition, the bank also wants to analyze reasons for customer attrition and what they can do to keep them
- The bank wants to build a data warehouse to support Marketing and other related customer care groups

How to Frame an Analytics Problem

Mini Case Study

Sample *Business* Problems	Qualifiers	Analytical Approach
• How can we improve on x? • What's happening real-time? Trends? • How can we use analytics differentiate ourselves • How can we use analytics to innovate? • How can we stay ahead of our biggest competitor?	Will the focus and scope of the problem change if the following dimensions change: • Time • People – how would x change this? • Risk – conservative/aggressive • Resources – none/unlimited • Size of Data?	Define an analytical approach, including key terms, metrics, and data needed.
Churn Prediction for Yoyodyne Bank Yoyodyne Bank How can we improve Net Present Value (NPV) and retention rate of the customers?	• **Time**: Trailing 5 months • **People**: Working team and business users from the Bank • **Risk**: the project will fail if we cannot determine valid predictors of churn • **Resources**: EDW, analytic sandbox, OLTP system • **Data**: Use 24 months for the training set, then analyze 5 months of historical data for those customers who churned	How do we identify churn/no churn for a customer? Pilot study followed full scale analytical model

Thanks